ChinaXiv合作期刊

知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

基于图数据库 Neo4j 的学者合作图谱分析

——以数字人文领域为例

熊回香 黄晓捷 陈子薇 李昕然
华中师范大学信息管理学院 武汉 430079

摘要：[目的 / 意义] 在深度数字化发展的背景下，数字人文成为跨学科深度融合的发展领域，学者之间的科研合作日益频繁，需要对其日趋复杂的合作关系进行分析与挖掘，帮助学者获得潜在的合作机会以促进学术交流。[方法 / 过程] 将学者、机构、关键词作为节点数据，合著、被引、任职、研究主题作为关系数据，构建学者合作图谱，基于图数据库 Neo4j 进行存储，并利用 Cypher 查询语言和 GDS 算法库对数字人文领域学者的合作社区发现、核心学者识别、合作趋势预测进行分析。[结果 / 结论] 实验结果证明，Neo4j 数据库较好地实现了数字人文领域学者合作网络的构建和图谱分析，能够帮助学者们在众多研究者当中快速地寻找与自己研究兴趣和方向高度关联的跨学科学者，从而促进数字人文领域学者合作与学科发展。

关键词：数字人文 学者合作 关系图谱 Neo4j

分类号：G250

引用格式：熊回香，黄晓捷，陈子薇，等. 基于图数据库 Neo4j 的学者合作图谱分析 —— 以数字人文领域为例 [J/OL]. 知识管理论坛，2022, 7(4): 465-260[引用日期]. http://www.kmf.ac.cn/p/308/.

数字人文作为计算机学科和人文学科交叉研究的一个跨学科领域，涉及的学科范围较广，包括语言学、文学、图书情报学和计算机科学等，由人文计算领域发展而来 [1]。在如今深度数字化时代，数字人文的研究热度越来越高，虽然我国学术界对其研究起步稍晚，但发展势头迅猛，获得了较好的发展前景 [2]。目前，我国数字人文的研究主要集中在对国外数字人文项目的调查与分析、利用数字化技术对人文艺术等资源进行可视化呈现及数字人文在图情档

基金项目：本文系国家社会科学基金年度项目“融合知识图谱与深度学习的在线学术资源挖掘推荐研究”（项目编号：19BTQ005）和中央高校基本科研业务费资助（创新资助项目）“个性化服务视角下数字档案馆用户画像模型构建研究”（项目编号：2020CXZZ124）研究成果之一。

作者简介：熊回香，教授，博士，博士生导师；黄晓捷，硕士研究生；陈子薇，硕士研究生；李昕然，硕士研究生，通信作者，E-mail：18730286901@163.com。

收稿日期：2022-03-23 发表日期：2022-08-26 本文责任编辑：刘远颖

领域的应用这三方面。此外，我国在数字人文的教育方面也取得了突破，上海图书馆、中国人民大学数字人文研究中心、武汉大学数字人文研究中心、北京大学信息管理系 KVision 实验室等科研机构深入推进数字人文和图情档的融合发展 [3]。在这样广阔的发展平台下，涌现出越来越多数字人文领域的学者，催生出庞大复杂的学术研究网络，主题多样，合作频繁。但是，如何在浩瀚无边的学术资源、学者、机构等信息中精准地找到自身需要的相关研究方向的合作对象是近些年科研合作预测研究的重点。因此，对学者合作关系网络进行分析，有利于发掘学者合作的规律和趋势，了解核心科研团队及研究主题，对把握此领域的发展状况具有重要意义，进而推动数字人文研究的发展和创新。

学者合作网络是相关领域学者在科研创作中因合著或被引关系而形成的复杂关联网络。学者合作网络可以加强学者之间的交流，对于知识共享、思维方式、科研创新等方面的进步有着不容小觑的作用。因此，目前越来越多的学者开始关注合作关系的研究，其中大多采用社会网络分析方法，刘培 [4]、刘志辉 [5]、邱均平 [6] 等学者基于社会网络分析法和关键词耦合分析法挖掘分析作者潜在的合作关系并构建合作网络。具体到数字人文领域，徐晨飞等运用文献信息统计分析工具以及社会网络分析方法对作者合著网络的网络结构特征、中心性、核心—边缘结构以及小型合著网络展开分析，总结该领域的科研合作特征 [7]；宫雪等通过高频关键词双聚类分析以及对合著网络和合著机构进行社会网络分析，从多角度探讨了当前国内数字人文研究的整体状况及研究热点 [8]。

近年来，开源或商用的图数据库不断涌现，主流的图数据库包括国内的 GDB[9]、Huge Graph[10] 以及国外的 Neo4j[11]、Tiger Graph[12] 等。这些图数据库集成了大量的社会网络分析方法与应用，主要包括中心性、路径查找、链接预测、社区检测和图可视化等，有助于发现知识图谱中的潜在知识，也能更好地发现社会网络中的合作关系 [13]。学术界内部分学者开始尝试使用图数据库开展社会网络分析研究。郭坤铭 [14] 利用 Neo4j 对异构网络中社会关系的分析优势，存储了百度百科上爬取的人物基本信息和关系，运用 Common Neighbors 算法进行网络结构相似度计算，并利用节点属性相似度预测所构建的异构网络中的人物社会关系。M. Kolomeets 等 [15] 利用图数据库 OrientDB 构建了 VKontakte 社交网络，使用 PageRank 评估了社交群体中最具影响力的意见领袖。丁洪丽 [16] 基于人员信息和话单等数据，采用 Neo4j 构建了多维关系网络并进行可视化，利用 Neo4j 中的查询分析功能挖掘人员关系，使得实验效率大幅提升。相较于传统的社会网络分析工具，图数据库能够展示大规模实体之间不断更新的庞大复杂关系，同时也能够使得网络节点和关系值间的查询更加简单快捷，在映射真实实体和关系方面具有天然优势 [17]。

针对数字人文领域中日益错综复杂的学术社交网络，如何对领域内的学者合作关系进行分析和挖掘逐渐成为该领域的一个研究重点。虽然传统的社会网络工具能够在一定程度上对学者合作网络进行分析，但对异构数据的处理仍有不足，且不具备图数据库的实时查询、预测推理、因果关系分析等功能 [13]。以 Neo4j 为主流的图数据库工具对多种关系数据的处理较为灵活，有望弥补这些不足。本文将在上述研究的基础上，运用 Neo4j 实现数字人文领域学者合作关系的构建与存储，并利用其强大的查询分析功能，快速便捷地查找相关学者并进行其合作关系的图谱分析，以期为相关领域的数字人文研究提供参考。

1 图数据库 Neo4j 及其应用优势

1.1 图数据库 Neo4j

随着互联网的不断发展，面对当下高并发的海量大数据和实时应用情景，图数据库以其易学、方便操作、高效处理复杂关系等独特的优势备受企业和学者的关注，它以图形数据结

ChinaXiv合作期刊
知识管理论坛，2022（4）:465-476
DOI: 10.13266/j.issn.2095-5472.2022.039

构存储实体及其相互关系，由节点、属性和边构成，其中节点表示数据实体，属性是节点的附属信息，边表示节点之间的关系，适合对关联关系复杂、动态关系多变的庞大数据进行存储和管理 [18]。与传统的关系型数据库相比，图数据库处理的是非结构化和不可预知的数据，更符合现在数据爆炸式增长与用户个性化需求的特点，并且有效支持实体间的关联关系，当加入新标签及新关系时，不需要调整先前的结构，拥有多层关联、最短路径、集中度测量等多种扩展功能，在社交网络、推荐系统、关系图谱等场景应用广泛，是大数据时代的新利器。

常见的图数据库有Neo4j、Flock DB、Graph DB、AllegroGrap等类型，其中，开源的Neo4j以其高性能、高稳定性、可扩展性强等优势成为当前应用最为广泛的原生图数据库之一 [19]。它采用原生图存储和处理数据，反映了关系网络中实体联系的本质，在查询中能以快捷的路径返回关联数据，表现出非常高效的查询性能；支持非结构化数据的存储与大规模数据的增长，能很好地适应需求的变化，具有很大的灵活性。此外，它还可以对实体间复杂的关系进行分析与推理，支持逻辑语言分析与面向约束的推理。Neo4j拥有自己的查询语言——Cypher语言，它是一种面向图分析、声明式、表达能力强的描述性图形查询语言 [20]，对用户十分友好，操作简便，主要使用的关键字有create（主要用于创建图形节点、关系及属性）、match（在已有图形数据库中匹配目标信息）、where（是match功能的条件）、return（完成匹配后，返回指定值），基于这些查询语句实现对图形数据的分析与推理。

1.2 Neo4j 分析学者合作网络的优势

随着网络技术的快速发展以及跨学科研究的日益突出，学者之间的合作关系也呈现复杂多样的特点，产生了越来越多的非结构化关联网络数据，Neo4j图数据库正是一个能够适应异构数据大规模增长和需求不断变化的数据库，它没有模式结构的定义，使用非结构化的方式来存储关联数据，不但适应能力强，而且自始至终都可以保持高效的查询性能，因此在处理学者之间复杂关系时显现出了独特的优势。

1.2.1 反映学者之间复杂的合作关系

合作关系是指学者们在学术研究过程中所进行的合作行为。常见的学者合作关系包括合著关系和引用关系。在学术网络中，如果两个学者的合著行为越频繁，那么他们更有可能兴趣相似且彼此信任，除此之外，学者的合著者也会与其他学者产生合著行为，基于这种学者间的合作关系便构建了学者合著网络，这种关系可以采用图结构存储，在此基础上，可以采取社会网络分析法和图挖掘算法对学者间的关系进行分析与聚类，从而发现最为匹配的合作者及合作团队。另外，学者间的另一种合作关系为引用关系，其被分为引用与被引，基于这两种引用行为，学者间构成了引文网络，是施引文献与被引成果的纽带，反映了引用者的借鉴、肯定以及相关问题的深层次研究。通常根据这样的引用关系实现资源聚合与学者聚合，以学者为节点，以文献之间的引用关系作为节点之间的联系边，以此构建相关引用文献之间的引用网络，从而更好地从引文关系网络中挖掘出核心学者或核心团队。不管是哪种合作关系，随着相关问题研究的多元化，学者间的合作关系也越来越复杂，而Neo4j恰好可以存储并反映这种量大、复杂而又变化的关联数据，支持大规模数据的增长与更新，且可清晰呈现各节点之间的关联关系。

1.2.2 实时查询目标学者的合作关系

除了存储功能，图数据库Neo4j的检索功能也非常强大，这依赖于Cypher查询语言，它是一种声明式图数据库查询语言，用法简洁且表现力丰富，查询效率高，拥有良好的扩展性，用户可以定制自己的查询方式。在检索功能中，Cypher语言由start、match、where、return 4个部分组成：① start表示在图中指定一个或多个起始节点，通过索引查找获得，也可以通过节点的编号直接获得；② match用于图形的匹配模

式，也是进行实例具体化的重要部分；③ where 提供过滤模式匹配结果的条件；④ return 用来指明在已经匹配查询的数据中，哪些节点、关系和属性是需要返回给客户端的。通过这样遍历查找的过程，容易定位聚焦到想要了解的学者节点，再利用条件的匹配，得到目标学者的合作关系，从而进行针对性分析。此外，Neo4j 还支持实时更新图数据库，且不影响已有的数据结构，这样可以不断地扩充现有关系图谱，展示越来越完备复杂的合作关系网络。

1.2.3 预测学者之间潜在的合作趋势

目前人物关系推理的方法主要有两种：基于本体的方法和基于图数据库的方法[21]。基于本体的人物关系推理时间复杂度较高，推理速度随人物关系数据量的增多而迅速降低，难以满足大数据时代下的人物关系推理需求，而基于图数据库的人物关系推理是人物关系数据分析的新趋势。图数据库的数据存储结构和数据查询方式都以图论为基础，适用于含有大量联系的人物关系数据的增删查改（CRUD）。基于图数据库的人物关系推理方法，首先将人物关系数据转换为图数据库的存储方式，然后采用图数据库查询语言进行人物关系分析[22]。作为支持效率高、扩展性强的声明式图查询语言及具有丰富开发模式的图数据库系统，Neo4j 存储学者关系知识图谱具有不可比拟的优势，复杂的关系链接也使其具备了推理能力，从而预测学者潜在的合作趋势，为不同领域、不同学科的科研合作提供可能的研究方向。

2 基于图数据库 Neo4j 的学者合作关系图谱构建

2.1 数据的选择与获取

本文选取中国知网学术资源总库中的 CSSCI 期刊作为数据来源进行数据获取，以“数字人文”或“人文计算”为主题进行检索，截至 2021 年 4 月 3 日，共检索到 615 篇文献。通过 NoteExpress 文献管理器对数据进行预处理，删除重复文献、会议征文、与数字人文主题不太相关的文献，最终获得有效文献 334 篇。对于多位作者署名的文献，本文统一选取前三位作者作为研究对象，经过重复项去除后，获得 410 个学者节点，244 个机构节点和 636 个关键词节点，数据处理结果示例见图 1；然后利用 Python 获取学者与学者之间的合著、被引关系，学者与机构之间的工作关系和学者与关键词之间的研究主题关系数据，本文主要基于上述 3 种节点和 4 种关系对学者合作关系进行图谱构建，数据模型见图 2。

Author	Organ	Title	Keywords	Journal
刘中华;焦基鹏;	上海工艺美术职业学	协同视域下图书馆耦合非遗数字人文服务模式研究	协同;;图书馆;;非遗;;数字人文	图书馆工作与研究
赵宇翔;练靖雯;	南京理工大学经济管	数字人文视域下文化遗产众包研究综述	文化遗产众包;;文化遗产机构;;文化	数据分析与知识发现
左娜;张卫东;	吉林大学管理学院;	数字人文多主体共生分析框架及其关键问题	数字人文;;多主体;;共生理论;;跨学	情报理论与实践
李天;	厦门大学中文系;	数字人文背景下的文学研究——量化方法在中西文	量化分析;;数字人文;;人文计算	厦门大学学报(哲学社会科学版)
冉从敬;何梦婷;黄海瑛;	武汉大学信息管理学	数字人文视阈下的比较文学可视化研究	数字人文;;比较文学;;可视化	厦门大学学报(哲学社会科学版)
邵晓宁;叶鹰;	南京大学信息管理学	国内外数字学术类研究的高引论文特征简析——兼	数字学术;;数字人文;;人文计算;;高	情报理论与实践
牛力;刘慧琳;曾静怡;	中国人民大学数据工	档案工作参与数字人文建设的模式分析	数字人文;;档案工作;;数字技术;;参	档案学通讯
张轩慧;赵宇翔;刘炜;朱庆	南京大学信息管理学	数字人文众包抄录平台用户体验优化的行动研究:基	数字人文;;众包抄录平台;;用户体验	中国图书馆学报
邓君;王阮;钟楚依;张子如	吉林大学管理学院;	数字人文视域下部分国家口述历史项目实践及启示	数字人文;;口述历史;;项目实践	图书情报工作
李道新;	北京大学艺术学院;	数字人文、影人年谱与电影研究新路径	数字人文;;影人年谱;;电影研究;;算	电影艺术
王兆鹏;邵大为;	中南民族大学文学与	数字人文在古代文学研究中的初步实践及学术意义	数字人文;;唐宋文学;;编年地图;;结	中国社会科学
[illegible]	北京航空航天大学外	数字人文视域下《论语》英译本的修辞对比研究	《论语》;;英译;;修辞;;数字人文	外语研究

图 1 数据处理结果示例

2.2 数据文件的导入

图数据要具体存储到图数据库中，就涉及到了特定的图数据模型，即关于采用什么实现方式来存图数据的问题。常见的图数据模型有属性图、超图和三元组。由于属性图模型直观且易于理解，能够描述绝大部分图的使用场景，Neo4j 采用的便是当下最流行的属性图模型。首先，将节点和关系数据的 Excel 文件都另存为“.csv”文件；然后利用 Cypher 语言的 create 语句，将节点文件和关系文件按照代码示例，见图 3，输入到代码编辑区；最后运行结果见图 4，清晰地展示了节点的个数、关系的对数以及学者合作关系图谱。具体于某一节点，以中国社会科学院文学研究所为例，通过此节点可查询到在

知识管理论坛，2022（4）:465-476
DOI: 10.13266/j.issn.2095-5472.2022.039

这个机构工作的两位学者，进而其合作的学者、研究主题等相关关系得到清晰的呈现，见图 5。

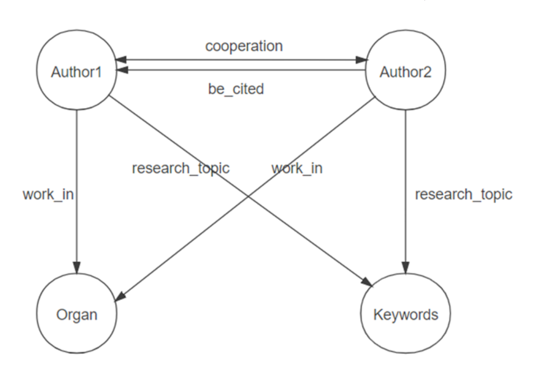


图 2 学者合作关系图谱数据模型

```
// 导入节点（以学者为例）
LOAD CSV WITH HEADERS FROM ‘file:///author.csv’ AS line
CREATE (s: author {Author: line.Author})
// 导入关系（以学者-学者的合著关系为例）
LOAD CSV WITH HEADERS FROM ‘file:///au-co.csv’ AS line
MATCH(from: author {Author: line. from_author}),
(to: author{Author: line.to_author})
MERGE (from)-[r: cooperation]→(to)
// 查询所有节点及关系
match (s) return s
```

图 3 导入数据代码示例

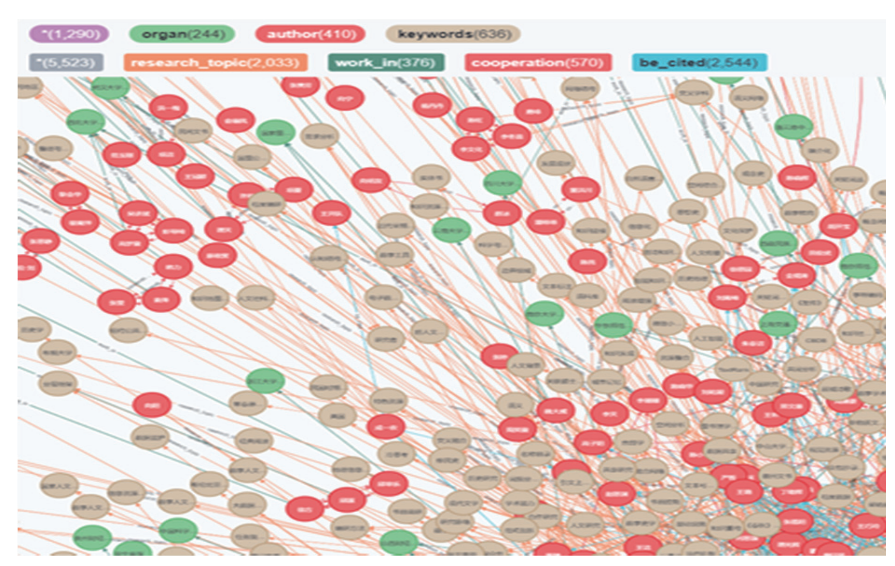


图 4 学者合作关系图谱构建样例

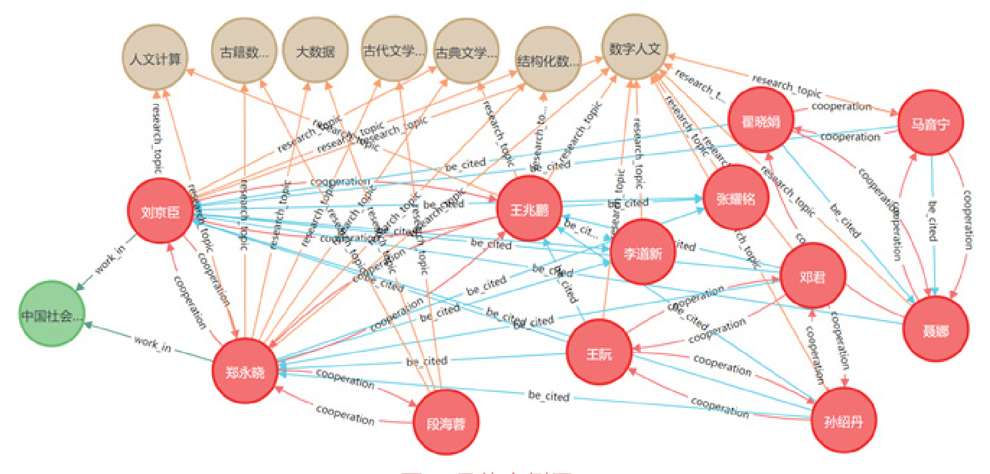


图 5 具体实例展示

3 基于图数据库 Neo4j 的学者合作图谱分析

面对庞大复杂的非结构化关系数据，图数据库 Neo4j 为技术的应用提供了有效的解决途径，但是通过梳理国内相关文献可知，目前利用 Neo4j 的内嵌图算法和 Cypher 查询语言进行数据分析与处理的研究相对较少，本文将充分利用 Neo4j 强大的图算法功能这一优势，对数字人文研究领域的学者合作网络进行分析。Neo4j 的算法库 Graph Data Science（GDS）可以实现各种复杂的社会网络分析，包括 centrality algorithms（中心性算法）、community detection algorithms（社区检测算法）、path finding algorithms（路径查找算法）、link prediction algorithms（链路预测算法）等。本文通过采用相关图算法，实现学者合作社区的发现、核心学者的识别以及学者合作趋势的预测，从不同角度为数字人文领域学者寻找自己的合作对象和资源提供借鉴。

3.1 合作社区发现

近年来，数字人文技术快速发展，吸引了越来越多的学者对相关问题进行广泛而深入的研究，因而构成了复杂的学者网络，社区结构便是复杂网络中的一个重要性质，体现为社区中的节点紧密相连且不同社区的节点稀疏连接[23]。它可以对有相似特征或共同属性的学者进行聚类，帮助学者发现并找到具有相似兴趣的同行或可以相互交流的跨学科合作者。在 Louvain、Label Propagation、infomap 等社区检测算法中，Louvain 在效率和效果上都表现较好，并能够发现层次性的社区结构。郭理等[24]使用经典数据集 American College Football 对 Louvain 算法与常用重叠社区发现算法 CPM、LFM 和 COPRA 进行实验对比，结果表明 Louvain 算法明显优于其他的算法。G. Drakopoulos 等[25]针对 Twitter 上的社交信息，在 Neo4j 中构建了争议性话题和普通性话题两个社交网络图，分别使用 Lonvain、Edge Betweeness、Waltrap 以及 CNM 等 4 种社区发现算法进行评估，实证发现 Louvain 算法产生的社区聚集性较高，社区成员的联系最为紧密。因此，本文选用 Louvain 方法在已构建学者合作网络中检测社区以实现对学者的模块化聚类，从而更好地分析学者聚集分区的特点以及它们加强或分散的趋势。在 GDS 中应用Louvain算法共发现100个学者合作社区，部分结果见图 6，按社区规模降序呈现。其中最大的社区包含 26 个学者，学者邓君、王阮、钟楚依、宋先智和孙绍丹之间合著频率较高，他们就数字人文视角下的历史项目进行分析研究；贺晨芝和徐孝娟对图书馆数字人文众包项目进行实践研究；李道新从电影艺术的角度分析了数字人文的应用路径等。由此可见，在模块化的社区里有合著频次较高的学者，也有跨学科相互引用的学者，同一社区的学者关联紧密程度较高，他们有着相通的研究方向和研究热点，表现出高度相似性。与此同时，图 7 的学者合作关系图谱也清晰地展现了不同社区学者的分布及其紧密程度，相同颜色的节点代表其处于同一个社区，研究主题相似的同时不同学者之间相互引证，进一步加强了学者之间的关联程度，为知识的交流与共享提供学习平台。

3.2 核心学者识别

核心学者是指在某个研究领域内研究成果数量较多、学术影响力较大、为该领域发展做出贡献的学者，他们是推动该领域学术进步的中坚力量[26]。核心学者的分析为学者们开展研究提供便利，帮助其全面地查询到自己感兴趣的核心学者群并快速查阅到该领域的核心科技文献，从而快速了解该领域研究的现状与不足，为自己深入研究奠定坚实的基础。中介中心性（Betweenness Centrality）算法是网络中心性衡量的经典指标，本文利用 GDS 中的 Betweenness Centrality 算法来衡量学者网络中不同节点的重要性，即检测其中一个节点对图中信息流的影响程度。该算法计算一个网络中所有节点对之间的未加权最短路径，每个节点根据通过该节点的最短路径的数量得到一个分数，更频繁地位于其他节点之间最短路径上的节点的得分更高。

ChinaXiv合作期刊
知识管理论坛，2022（4）:465-476
DOI: 10.13266/j.issn.2095-5472.2022.039



Size	Nodes
26	邓君 王阮 钟慧依 李道新 王兆鹏 邵大为 陈静 曾泽南 姚建华 贺晨芝 徐孝娟 孙绍丹 郑永晓 段海蓉 韩翰哲 张久珍 宋先智 刘京臣 孙辉 侯莹 赵洪雅 柯平 宫平 张晓梅 徐彤阳 杨明睿
24	刘中华 焦慧鹏 李慧楠 王晓光 陈涛 单蓉蓉 高建辉 王军 王贵海 许继新 王新雨 陈雪蓬 杨佳颖 严承希 周晨 胡爱民 曾蕾 范炜 叶焕辉 朱娜 周文杰 赵小兵 丁绪辉 傅柱
24	邵晓宁 叶鹰 段力萌 魏凤 欧阳剑 彭松林 李臻 赵文娟 蒋萌 黄水清 王若宸 母睦然 朱学芳 胡以涛 惠富平 唐乐 安娃 朱本军 聂华 尚晓倩 谷学强 刘浏 严哲 孟凯
20	赵宇翔 练靖雯 张轩慧 刘炜 刘周颖 张磊 韩文婷 宋士杰 王日芬 王薇龙 刘洪 汪莉 宋小康 饶俊丽 谢蓉 聂娜 岑旻莲 翟晓娟 马育宁 曾辉
18	许鑫 陆柳梦 邓诺萝 朱慧敏 杨沉 郑泽泉 蔡迎春 张玲 薛志红 黄钰新 王远智 周亚 崔春 毕强 郭金龙 李国刚 牛赞宇 刘燕权
17	唐燕 刘小楷 李健 田燕飞 盛小平 杨鹤林 李宾 朱前东 于亚秀 吴加琪 董梅香 赵子非 杨滋荣 熊回香 蒋合领 苏芳荔 常人杰
16	左娜 张卫东 李子林 龙家庆 王玉珏 张斌 杨茜茜 霍艳芳 何思源 魏永忠 周建新 阿德莱德·赫夫伯格 房小可 谭富强 谢永宪 王巧玲

图 6 学者合作社区发现部分结果

图 7 学者合作社区部分关系图谱

在 GDS 中，Betweenness Centrality 算法通过对 410 位学者的最短路径进行打分，按照分数降序排列的同时给每位学者赋予一个编号，识别结果见表 1。学者刘炜得分最高，赵宇翔次之。得分越高，说明这些学者在数字人文研究领域的活跃度较高，同时也说明他们在此领域建树颇丰并有着较高的学术影响力。根据识别结果数据绘制散点图，如图 8 所示，在节点 16

后出现了明显的断崖式下降，由此初步认为前16 位学者可被识别为数字人文领域研究的核心学者，在这些核心学者中，刘炜和夏翠娟工作于上海图书馆，朱学芳和叶鹰工作于南京大学，赵宇翔工作于南京理工大学，王晓光工作于武汉大学等，从一定程度上可以反映出这些学者的工作单位是其科学研究的主要阵地，以他们为代表拥有着该领域研究的核心团队，他们带领自己的学生及合作者深入地开展着数字人文的研究，成果颇多。其中，上海图书馆主持有关于数字人文的国家哲学社会科学基金项目，夏翠娟和刘炜学者是数字人文团队中的重要成员，其团队基于数字人文构建了家谱知识服务平台 [27]、名人手稿档案库 [28]、中文古籍联合目录及循证平台 [29] 等，在国内将数字人文的研究和应用推向新的发展阶段。为了进一步清晰地反映核心学者，可利用 Neo4j 所呈现的图谱中学者节点的大小来反映其在数字人文研究领域中所处的位置，如图 9 所示，节点越大，其学术影响力越大。这对于相关研究者找寻领域内核心学者具有重要参考意义，且更加方便快捷，清晰明了。

表 1 部分核心学者识别结果

Order（序号）	Node（节点）	Score（得分）	Order（序号）	Node（节点）	Score（得分）
1	刘炜	11 516.430	10	左娜	4 384.182
2	赵宇翔	11 374.860	11	黄水清	3 704.112
3	朱学芳	9 463.338	12	赵生辉	2 982.959
4	欧阳剑	8 295.446	13	叶鹰	2 904.088
5	夏翠娟	5 082.038	14	赵薇	2 820.388
6	邓君	4 649.169	15	卢章平	2 734.507
7	张卫东	4 522.494	16	王晓光	2 212.367
8	牛力	4 433.821	……	……	……
9	许鑫	4 433.689	410	刘慧平	0

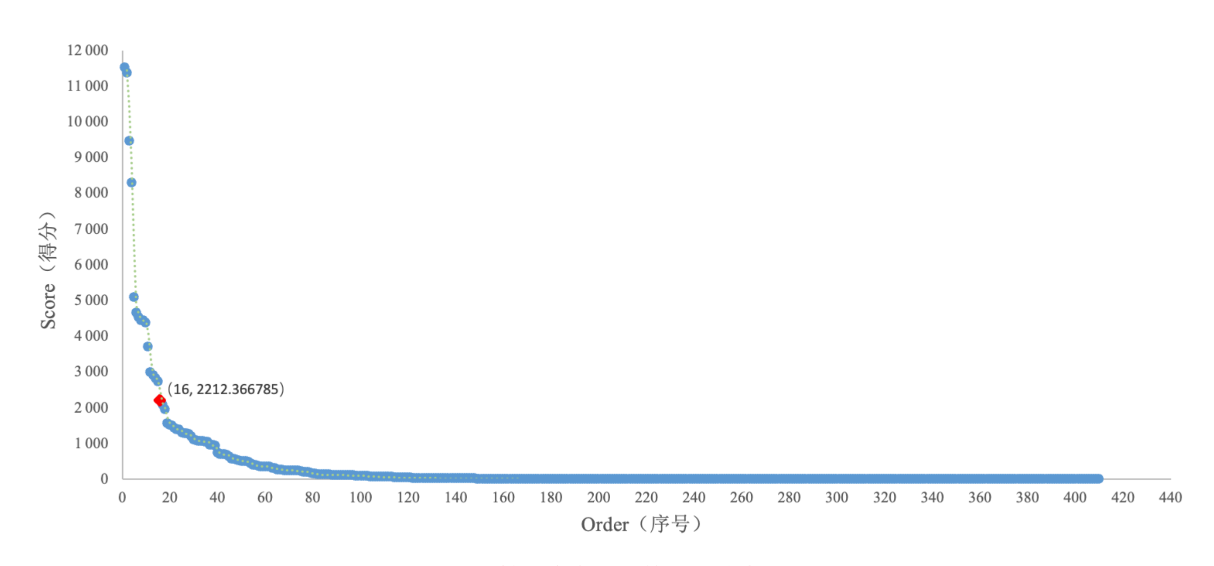


图 8 核心学者识别的得分散点图

知识管理论坛，2022（4）:465-476
DOI: 10.13266/j.issn.2095-5472.2022.039

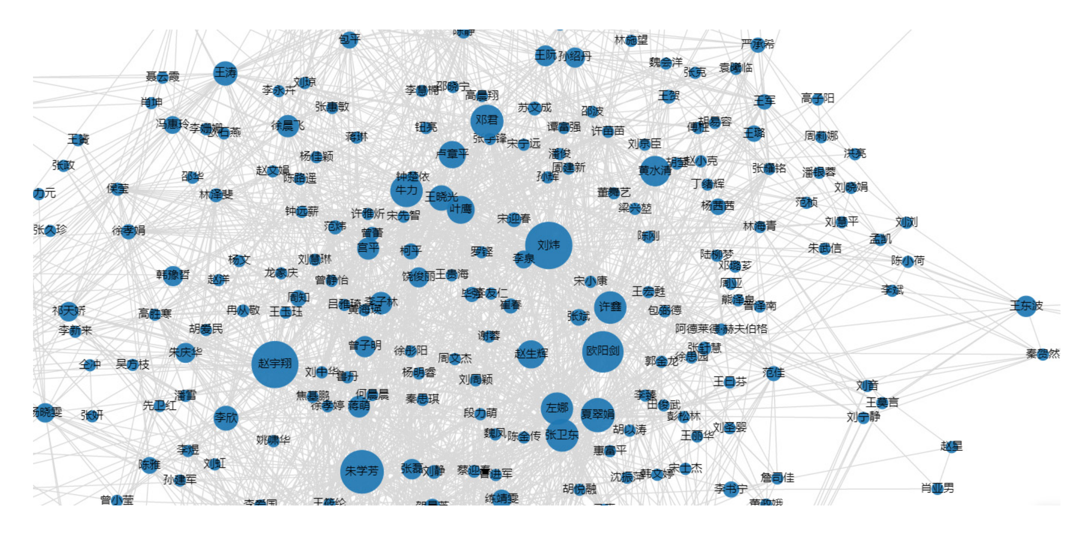


图 9 部分核心学者关系图谱

3.3 合作趋势预测

在大数据时代，学术研究的合作化趋势日益明显，作为科研活动的重要组成部分，合作形式在提升科研效率、促进科研产出时发挥着极其重要的作用。研究表明，在过去的 20 多年里，各个学科中的合作研究的数量都呈显著增长趋势，具有相同研究领域、相似研究方向的学者更易于在未来进行合作[30]。但是，由于时间、空间位置的阻碍，学者们很难在浩如烟海的学者群体里准确找到与自身研究方向相近的学者，分析挖掘学者潜在的合作对象可以有效提高其科研效率。本文利用 GDS 中的链路预测算法对节点之间的接近度进行计算，从而帮助学者找到潜在的合作机会。

链路预测算法是指通过已知节点的特征信息以及网络拓扑结构，预测尚未产生连接的节点对之间出现连边的可能性。常见的链路预测算法包括基于邻居节点的链路预测以及基于共有邻居的链路预测，其中基于邻居节点的算法包括所有邻居（total neighbors）以及连接偏好（preferential attachment）等，基于共有邻居的算法包括共有邻居（common neighbors）、资源优化（resource allocation）以及 AA（adamic adar）算法等[31]。D. Liben-Nowell 等[32]、T. Zhou 等[33]通过实验对多种链路预测算法对比分析发现 AA 算法效果相对较优。AA 算法基于共有邻居的相邻节点集合，并对集合数量进行非线性归一化处理，计算两个节点的紧密度，其预测网络中学者合作链接的公式如下所示：

$$A(x,y)=\sum_{u\in N(x)\cap N(y)}\frac{1}{\log|N(u)|} \qquad \text{公式（1）}$$

在该公式中，当计算结果的值为 0 时，表示两个节点不靠近；当值越大时则表示节点越靠近。

在上述学者合作社区发现分析中，相较于不同社区来说，同一社区学者的合作关系更为紧密，但是尽管在同一社区，他们的合作也存在疏密之分，本文选取第四大学者合作社区，以核心学者“刘炜”为研究对象，利用上述公式和 Cypher 查询语言“MATCH (s1:author{Author:‘刘炜’}),MATCH (s2:author{Author:‘*’}),RETURN gds. alpha. linkprediction. adamicAdar (s1, s2) AS score”计算并呈现刘炜与其同一社区中其他学者的可能链接程度，预测值分数见表 2。其中刘炜和赵宇翔可能产生链接关系的得分最高，说明他们发生合作的可能性最大，而刘炜和汪莉进行合作的可能性则最小。与此同时，通过 Cypher 查询语句将刘炜所在的社区的学者合作关系图谱进行呈现，见图 10。这个图表明了同一社区的学者关联紧密，但其中也存在少部分

学者之间未建立直接的合作关系，如刘炜与岑灵莲、曾辉、刘洪、汪莉这 4 位学者，相对应他们的合作链接预测值也较低。通过分析表 2 和图 10 不难发现，在已产生直接连接的学者中，宋士杰得分最低，此分数可确定为产生新链接的最低阈值，即当未发生直接连接的两个学者得分大于这个阈值时，则能说明其更能产生链接，其合作的可能性更大。由此可以看出刘炜与岑灵莲、曾辉、刘洪更能进行有效的科研交流，合作趋势较为明显。

综上所述，Neo4j 的语句查询和算法分析功能是学者合作趋势预测的有效工具，为学者寻找自己的合作伙伴节省时间，提高合作效益。在学者交流活动日趋频繁的背景下，科研合作已然成为学者推动学术研究发展的必要形式，学者间的合作越多样多元，那么该领域的学术交流氛围越活跃高效，不同的思维碰撞推动数字人文领域的多元化、跨学科式发展。

表 2 同一社区学者之间合作预测值得分表

目标学者	同一社区其他学者	得分
刘炜	赵宇翔	21.794
	张磊	10.503
	张轩慧	8.044
	练靖雯	6.496
	刘周颖	4.960
	谢蓉	4.395
	饶俊丽	3.330
	刘洪	3.243
	聂娜	3.143
	翟晓娟	3.140
	马音宁	3.138
	王筱纶	3.059
	宋小康	2.993
	岑灵莲	2.695
	曾辉	2.695
	王曰芬	2.063
	韩文婷	1.731
	宋士杰	1.731
	汪莉	0.909

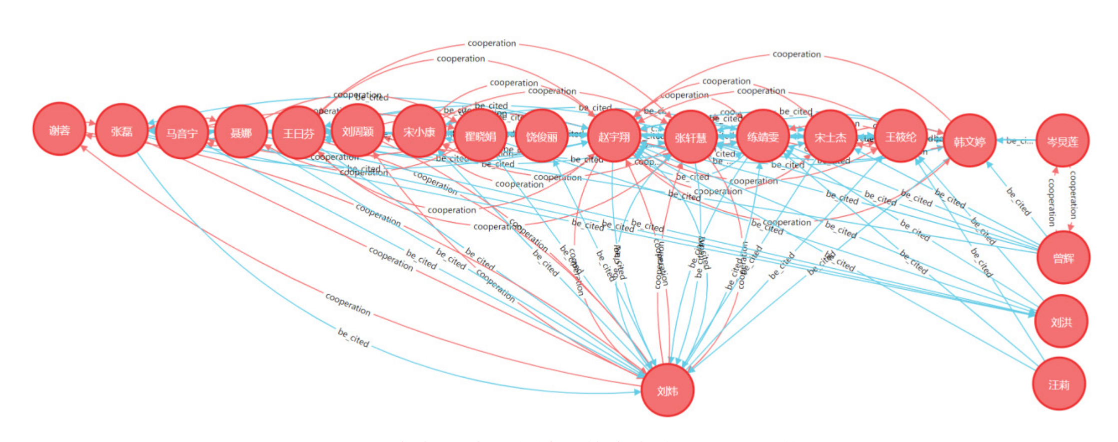


图 10 学者刘炜所在社区的学者合作关系图谱

4 结语

随着数字时代的深入发展，“数字人文”对实施文献抢救性保护、提供公共文化服务、弘扬中华民族优秀传统文化等方面都具有重要的现实意义。在我国，数字人文作为专业学术研究已开始加速发展，而且由这种跨学科的研究范式孕育而生的研究成果也将通过更多的合作形式来呈现。对于科研工作者来说，合作能够促使学者产生新的想法、新的研究思路，能够提高合作者的产出量和影响力；对于学科发展来说，合作能够促使新的知识体系的形成，开阔学者的知识视野和更新学者的知识结构，在帮助学者们快速高效地寻找与自己研究兴趣和方向高度关联的跨学科学者、加强交流合作的同时推动数字人文的多学科深度融合发展。

ChinaXiv合作期刊
知识管理论坛，2022（4）:465-476
DOI: 10.13266/j.issn.2095-5472.2022.039

本文利用处理复杂关联数据的利器——图数据库 Neo4j 对我国数字人文的研究主体（即学者）及其间关系进行存储分析，利用 GDS 算法库实现了学者合作社区的发现、核心学者的识别以及合作趋势的预测。虽然社会网络分析方法从中心性、凝聚子群、核心—边缘等不同角度在各种关联网络结构的分析中非常普遍，但是本文利用图数据库 Neo4j 实现了传统的社会网络分析方法能够达成的功能外，还实现了数据存储、实时更新、即查即得、预测推理等功能，这是对社会网络分析方法的有力补充，为社会网络分析提供了新的思路与方法。

此外，本文的不足之处在于：①在获取相关文献时忽略了一些篇名没有以“数字人文”或“人文计算”命名但研究内容为“数字人文”的研究成果，使得学者节点和关系数据量偏小，在完整性上稍有欠缺；②数据量越大，复杂度越高，图数据库 Neo4j 处理数据的优势就越明显，但本文在研究图数据库 Neo4j 的功能应用上较为简单，没有很好地发挥出其数据分析的优势。因此，在未来的研究中，笔者将继续深入学习 Neo4j 极其强大的数据分析功能，不断扩大更新学者的数据量，从而充分展现学者之间复杂的合作关系，为学者们进行潜在科研合作提供借鉴。

参考文献：

[1] 柯平 , 宫平 . 数字人文研究演化路径与热点领域分析 [J]. 中国图书馆学报 , 2016, 42(6): 13-30.

[2] 潘连根 . 数字人文在档案领域中应用的理性思考 [J]. 档案与建设 , 2020(7): 6-10.

[3] 牛力 , 高晨翔 , 张宇锋 , 等 . 发现、重构与故事化 : 数字人文视角下档案研究的路径与方法 [J]. 中国图书馆学报 , 2021, 47(1): 88-107.

[4] 刘蓓 , 袁毅 , BOUTIN E. 社会网络分析法在论文合作网中的应用研究 [J]. 情报学报 , 2008, 27(3): 407-417.

[5] 刘志辉 , 张志强 . 作者关键词耦合分析方法及实证研究 [J]. 情报学报 , 2010, 29(2): 268-275.

[6] 邱均平 , 刘国徽 . 基于社会网络和关键词分析的作者合作研究——以国内知识管理领域为例 [J]. 情报科学 , 2014, 32(6): 3-7, 13.

[7] 徐晨飞 , 赵文娟 . 我国数字人文研究领域作者合著网络分析 [J]. 图书馆论坛 , 2019, 39(11): 14-24.

[8] 宫雪 , 杨颖 . 我国数字人文研究热点及合著网络可视化分析 [J]. 图书情报导刊 , 2019, 4(6): 39-45.

[9] 图数据库 GDB- 帮助中心 - 阿里云 [EB/OL].[2021-10-10]. https: //help.aliyun.com/product/102714.html.

[10] HugeGraph [EB/OL].[2021-10-10].https: //hugegraph.github.io/hugegraph-doc/.

[11] Native Graph Database | Neo4j Graph Database Platform [EB/OL].[2021-10-10]. https: //neo4j.com/product/neo4j-graph-database/.

[12] Graph Database | Graph Analytics Platform | TigerGraph [EB/OL].[2021-10-10]. https: //www.tigergraph.com/.

[13] 刘春江 , 李姝影 , 胡汗林 , 方曙 . 图数据库在复杂网络分析中的研究与应用进展 [J]. 数据分析与知识发现 , 2022, 6(7): 1-11.

[14] 郭坤铭 . 基于异构网络的关系推理与预测方法研究 [D]. 太原 : 太原理工大学 , 2017.

[15] KOLOMEETS M, CHECHULIN A, KOTENKO I V. Social networks analysis by graph algorithms on the example of the VKontakte social network[J]. Journal of Wireless Mobile Networks, 2019, 10(2): 55-75.

[16] 丁洪丽 . 基于 Neo4j 图数据库的人员关系挖掘 [J]. 电讯技术 , 2020, 60(7): 771-777.

[17] CHU Z, YU J, HAMDULLA A, A novel deep learning method for query task execution time prediction in graph database[J]. Future generation computer systems, 2020, 112: 534-548.

[18] 李金阳 . 图数据库在图书馆的应用研究 [J]. 图书馆 , 2020(11): 109-115.

[19] FLEMING J, LEVY S, NAG P, et al. Graph database system and method for facilitating financial and corporate relationship analysis: US, US8674993 B1[P].2014.

[20] 张维冲 , 王芳 , 黄毅 . 基于图数据库的贵州省大数据政策知识建模研究 [J]. 数字图书馆论坛 , 2020(4): 30-38.

[21] 于娟 , 黄恒琪 , 席运江 , 等 . 基于图数据库的人物关系知识图谱推理方法研究 [J]. 情报科学 , 2019, 37(10): 8-12.

[22] DRAKOPOULOS G, KANAVOS A, MYLONAS P, et al. Defining and evaluating Twitter influence metrics: a higher-order approach in Neo4j[J].Social network analysis & mining, 2017, 7(1): 52.

[23] GIRVAN M, NEWMAN M E J. Community structure in social and biological networks[J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[24] 郭理，王嘉岐，张恒旭，等．基于 Louvain 重叠社区发现算法 [J]. 石河子大学学报 (自然科学版), 2020, 38(3): 384-389.
[25] DRAKOPOULOS G, GOURGARIS P, KANAVOS A. Graph communities in Neo4j[J]. Evolving systems, 2020, 11(3): 397-407.
[26] 王曰芬，杨雪，余厚强，等．人工智能科研团队的合作模式及其对比研究 [J]. 图书情报工作，2020, 64(20): 14-22.
[27] 夏翠娟，刘炜，陈涛，等．家谱关联数据服务平台的开发实践 [J]. 中国图书馆学报，2016(3): 27-38.
[28] 夏翠娟，张磊，贺晨芝．面向知识服务的图书馆数字人文项目建设：方法、流程与技术 [J]. 图书馆论坛，2018(1): 1-9.
[29] 夏翠娟，林海青，刘炜．面向循证实践的中文古籍数据模型研究与设计 [J]. 中国图书馆学报，2017(6): 16-34.
[30] GLANZEL W, SCHUBERT A. Analysing scientific networks through co-authorship[M]//Handbook of quantitative science and technology research. Dordrecht: Springer, 2004: 257-276.
[31] LU L, ZHOU T. Link prediction in complex networks: a survey[J]. Physica A: statistical mechanics and its applications, 2011, 390(6): 1150-1170.
[32] LIBEN - NOWELL D, KLEINBERG J. The link - prediction problem for social networks[J]. Journal of the American Society for Information Science and Technology, 2007, 58(7): 1019-1031.
[33] ZHOU T, LU L, ZHANG Y C. Predicting missing links via local information[J]. The European physical journal B, 2009, 71(4): 623-630.

作者贡献说明：
熊回香： 研究整体思路框架指定、论文指导；
黄晓捷： 数据收集与处理、论文撰写；
陈子薇： 数据处理、论文修改；
李昕然： 论文修改、最终版本修订。

Analysis of Scholar Collaboration Map Based on Graph Database Neo4j ——Taking the Field of Digital Humanities as an Example

Xiong Huixiang Huang Xiaojie Chen Ziwei Li Xinran
School of Information Management, Central China Normal University, Wuhan 430079

Abstract: [Purpose/Significance] In the context of deep digital development, digital humanities as a development field of interdisciplinary deep integration, the scientific research cooperation among scholars is becoming more and more frequent. It is necessary to analyze and excavate the increasingly complex cooperation relationship, to help scholars obtain potential cooperation opportunities to promote academic exchanges. **[Method/Process]** In this paper, scholars, institutions and keywords were used as node data, and coauthors, citations, posts and research topics were used as relational data to build scholar-collaboration graphs, which was stored based on the graph database Neo4j. Cypher query language and GDS algorithm library were used to analyze the cooperation community discovery, core scholar identification and cooperation trend prediction of scholars in the field of digital humanities. **[Results/Conclusion]** The experimental results show that Neo4j can better realize the construction and analysis of scholars' cooperation network in the field of digital humanities. It can help scholars quickly find interdisciplinary scholars who are highly related to their research interests and directions among many researchers, so as to promote scholars' cooperation and discipline development in the field of digital humanities.

Keywords: digital humanities   scholar cooperation   relationship map   Neo4j